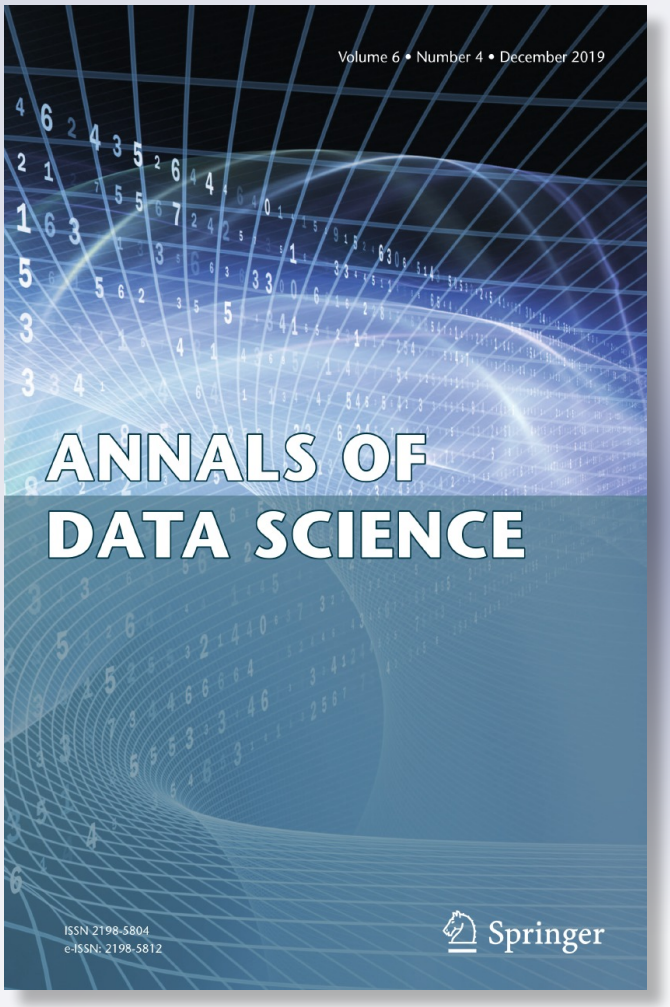
On Regularisation Methods for Analysis of High Dimensional Data

Tanin Sirimongkolkasem & Reza Drikvandi

Annals of Data Science

ISSN 2198-5804
Volume 6
Number 4

Ann. Data. Sci. (2019) 6:737-763
DOI 10.1007/s40745-019-00209-4



Springer

Annals of Data Science (2019) 6(4):737–763
https://doi.org/10.1007/s40745-019-00209-4

On Regularisation Methods for Analysis of High Dimensional Data

Tanin Sirimongkolkasem[1] · **Reza Drikvandi**[2]

Received: 14 January 2019 / Revised: 23 March 2019 / Accepted: 7 April 2019 /
Published online: 13 April 2019

Abstract
High dimensional data are rapidly growing in many domains due to the development of technological advances which helps collect data with a large number of variables to better understand a given phenomenon of interest. Particular examples appear in genomics, fMRI data analysis, large-scale healthcare analytics, text/image analysis and astronomy. In the last two decades regularisation approaches have become the methods of choice for analysing such high dimensional data. This paper aims to study the performance of regularisation methods, including the recently proposed method called de-biased lasso, for the analysis of high dimensional data under different sparse and non-sparse situations. Our investigation concerns prediction, parameter estimation and variable selection. We particularly study the effects of correlated variables, covariate location and effect size which have not been well investigated. We find that correlated data when associated with important variables improve those common regularisation methods in all aspects, and that the level of sparsity can be reflected not only from the number of important variables but also from their overall effect size and locations. The latter may be seen under a non-sparse data structure. We demonstrate that the de-biased lasso performs well especially in low dimensional data, however it still suffers from issues, such as multicollinearity and multiple hypothesis testing, similar to the classical regression methods.

Keywords De-biased lasso · High dimensional data · Lasso · Linear regression model · Regularisation · Sparsity

✉ Reza Drikvandi
r.drikvandi@mmu.ac.uk

1 Statistics Section, Department of Mathematics, Imperial College London, London, UK

2 Department of Computing and Mathematics, Manchester Metropolitan University, Manchester, UK

1 Introduction

1.1 Background and Importance

"High dimensional" refers to the situations where the number of covariates or predictors is much larger than the number of data points (i.e., $p \gg n$). Such situations happen in many domains nowadays where the rapid development of technological advances helps collect a large number of variables to better understand a given phenomenon of interest. Examples occur in genomics, fMRI data analysis, large-scale healthcare analytics, text/image analysis and astronomy, to name but a few.

In the last two decades regularisation approaches such as lasso, elastic net and ridge regression have become the methods of choice for analysing such high dimensional data. Much work has been done since the introduction of regularisation in tackling high dimensional linear regression problems. Regularisation methods especially lasso and ridge regression [10,31,40] have been applied to many applications in different disciplines [1,15,23,26]. The theory behind regularisation methods often relies on the sparsity assumptions to achieve theoretical guarantees in their performance, ideally when dealing with high dimensional data. The performance of regularisation methods has been studied by many researchers, however conditions other than sparsity, such as the effects of correlated variables, covariate location and effect size have not been well understood. We investigate this in high dimensional linear regression models under sparse and non-sparse situations.

In this paper, we consider the high dimensional linear regression model

$$y = X\beta + \varepsilon, \quad p \gg n, \tag{1}$$

where $y = (y_1, \ldots, y_n) \in \mathbb{R}^n$ is the response vector, $X \in \mathbb{R}^{n \times p}$ is the design matrix for covariates $x_1, \ldots, x_n$, the vector $\beta \in \mathbb{R}^p$ contains the unknown regression coefficients, and $\varepsilon \in \mathbb{R}^n$ is the random noise vector. We assume, without loss of generality, that the model does not have any intercept terms by mean-centring all the response and covariates.

We assume no prior knowledge on β. It is well-known that the ordinary least square (OLS) solution for estimating β is $\hat{\beta}^{\text{OLS}} = (X^T X)^{-1} X^T y$ [10]. However, when $p > n$, X is no longer full rank, and the OLS results in infinitely many solutions, leading to over-fitting in the high dimensional case [14]. This kind of ill-posed problems arises in many applications as discussed above. Regularisation methods that impose penalty on the number of unknown parameters β is therefore a general and popular way to overcome the issue of ill-posed problems.

Issues due to the curse of dimensionality become apparent in the case of $p \gg n$. A particular example occurs in fMRI image analysis, where selection from a large number of brain regions could lead to insensitive models on top of over-fitting [30]. Also, numerical results from utilising regularisation methods in high dimensional data are unsatisfactory in terms of identifying one that performs the best most of the time [7]. To tackle these issues, sparsity assumption utilises the idea of "less is more" [14], referring to the phenomenon that an underlying data structure can mostly be explained

by few out of many features. Such assumption would help some regularisation methods to at least achieve consistent variable selection even when $p \gg n$ [18].

Data structures are not necessarily sparse in real-world applications of high dimensional data, and sparsity assumptions are difficult to hold in practice [3]. However, regularisation methods are often applied to those applications despite their limitations. In this paper, we mainly aim to study the following problems which have not been well understood or never studied previously:

i. The effects of data correlation on the performance of common regularisation methods.
ii. The effects of covariate location on the performance of common regularisation methods.
iii. The impact of effect size on the performance of common regularisation methods.
iv. The performance of the recently developed de-biased lasso [33,37] in comparison to the common regularisation methods.

In our investigations we evaluate the performance of the regularisation methods by focusing on their variable selection, parameter estimation and prediction performances under the above situations. We also use simulations to explain the curse of dimensionality with a fixed-effect Gaussian design.

1.2 Related Work

Lasso and ridge regression, which use L_1 and L_2 penalties respectively (see Sect. 2.1), are the two most common regularisation methods. Many novel methods have been built upon them. For example, Zou and Hastie [40] developed the elastic net that uses a combination of these two penalties. The elastic net is particularly effective in tackling multicollinearity, and it can generally outperform both lasso and ridge regression under such situation. The study on elastic net had relatively low dimensions with the sample size larger than the number of covariates [40]. Moreover, the number of covariates associated with truly non-zero coefficients was smaller than the sample size. Studies with similar kinds of settings were also used when developing other novel methods [25, 36,39]. Other new approaches with variations of standard techniques have also been investigated [13,22,26]. Also, reducing bias of estimators such as the lasso estimator is recently used to tackle issues of 0 standard errors and biased estimates [2,37]. Beforehand, a method called the bias-corrected ridge regression utilised the idea of bias reduction by projecting each feature column to the space of their compliment columns to achieve, with Gaussian noise, asymptotic normality for a projected ridge regression estimator under a fixed design [4]. Regularisation methods were also evaluated in other situations with classification purposes [1,12,13,23,24,26,35].

Statistical inference such as hypothesis testing with regularisation methods was difficult for a long time due to the mathematical limitations and the highly biased estimators in high dimensional models. Obenchain [27] argue that inference with biased estimators could be misleading when they are far away from their least squares region. The asymptotic theory has also shown that the lasso estimates can be 0 when the true values are indeed 0 [21], which can explain why the bootstrap with lasso estimators can lead to a 0 standard error [31]. Park and Casella [28] developed a

Bayesian approach to construct confidence intervals for lasso estimates as well as its hyperparameters. They considered a Laplace prior for the unknown parameters β in the regression model (1) conditional on unknown variance of an independent and identically distributed Gaussian noise, leading to conditional normality for y and β. However, they did not account for the presence of bias in parameter estimators when using regularisation methods. The recent de-biased lasso (see Sect. 2.2) instead reduces the bias of lasso and enables to make statistical inference about a low dimensional parameter in the regression model (1). It is unknown whether or not the de-biased lasso can outperform the lasso and other regularisation methods when increasing the data dimension in sparse and non-sparse situations.

More recently, [6] conducted a theoretical study on the prediction performance of the lasso. Their main finding was that the incorporation of a correlation measure into the tuning parameter could lead to a nearly optimal prediction performance of the lasso. Also, [29] proposed the spike-and-slab lasso procedure for variable selection and parameter estimation in linear regression.

2 Regularisation Methods in High Dimensional Regression

2.1 Regularisation with a More General Penalty

Given the high dimensional linear regression (1), the regularisation with L_q penalty minimises

$$\frac{1}{n}\sum_{i=1}^{n}(y_i - x_i\beta)^2 + \lambda\Big(\sum_{j=1}^{p}\beta_j^q\Big)^{1/q}, \tag{2}$$

where the first term is the squared error loss from the OLS, and the second term is a general L_q penalty on regression coefficients with $\lambda \geq 0$ being a tuning parameter for controlling the amount of shrinkage.

Two special cases of (2) are the lasso with L_1 penalty (i.e., $q = 1$) and the ridge regression with L_2 penalty (i.e., $q = 2$). Also, the subset selection emerges as $q \to 0$, and the lasso uses the smallest value of q (i.e., closest to subset selection) that yields a convex problem. Convexity is very beneficial for computational purposes [32].

Since the lasso provides a sparse solution (i.e., the number of non-zero parameter estimates are smaller than the sample size n) [31,32], lasso regression requires the sparsity assumption, that is, many of the covariates are assumed to be unrelated to the response variable. It is appealing to be able to identify, out of a large number of predictors, a handful of them that are main contributions to some desired predictions, particularly in genome-wide association studies (GWAS) [2,5,35]. This leads to parsimonious models from which the selected variables can be further examined, as well as greatly reducing subsequent computational costs in predictions.

A limitation of the lasso is that when there are contributing variables in a correlated group, lasso tends to select only few of them in a random manner. Yuan and Lin [36] proposed the group lasso method for performing variable selection on groups

of variables to overcome the issue, given prior knowledge of the underlying data structure. Also, the choice of λ in lasso may not satisfy the oracle properties, which can lead to inconsistent selection results in high dimensions [9,39]. As discussed in the introduction, Zou [39] developed the adaptive lasso to allow weighted L_1 penalty on individual coefficients, and showed that the new penalty satisfies the oracle properties. Further issues arise with high dimensional data, which can be summarised as curse of dimensionality. Roughly speaking, curse of dimensionality is a phenomenon at which ordinary approaches (in this case regularisation approaches) to a statistical problem are no longer reliable when the associated dimension is drastically high. We particularly investigate this in Sect. 3.

2.2 The De-Biased Lasso

De-biased lasso [33,37] is a lasso-based method that aims to reduce the bias of the lasso estimator. Also, unlike the original lasso, the de-biased lasso enables us to make statistical inferences, for example, to conduct component-wise hypothesis testing in high dimensional models [33]. It is known that the lasso estimator $\hat{\beta}^{\text{lasso}}$ fulfils the Karush–Kuhn–Tucker (KKT) conditions: [33,37]

$$-\frac{1}{n}X^T(y - X\hat{\beta}^{\text{lasso}}) + \lambda(\partial||\hat{\beta}^{\text{lasso}}||_1) = \underline{0}_p, \tag{3}$$

where $\underline{0}_p \in \mathbb{R}^p$ is the zero vector and $\partial||\beta||_1$ denotes the sub-differential of the l_1-norm of β with

$$\begin{aligned}(\partial||\beta||_1)_j &= \text{sign}(\beta_j), \quad \beta_j \neq 0,\\ (\partial||\beta||_1)_j &\in [-1, 1], \quad \beta_j = 0.\end{aligned}$$

The sub-differential at $\beta_j = 0$ for any $j = 1, 2, \ldots, p$ is a convex set of all possible sub-gradients since the l_1 norm is not differentiable at that point.

Substituting $y = X\beta + \varepsilon$ and $G = \frac{1}{n}X^TX$ into Eq. (3) and using the strong sparsity assumption, we get the following estimator [33]

$$\hat{\beta}^{\text{debiased}} = \hat{\beta}^{\text{lasso}} + \frac{1}{n}\widetilde{G}X^T(y - X\hat{\beta}^{\text{lasso}}) + R, \tag{4}$$

where $\widetilde{G}$ is an inverse matrix approximation of G, and R is a residual term from using an approximated inverse matrix [33, for details see]. One can view the second term on the right side of (4) as a bias correction to the lasso estimator [37]. Together with linear Gaussian setting and some restrictions in choosing an optimal λ, $\hat{\beta}^{\text{debiased}}$ yields a statistic that asymptotically follows a multivariate Gaussian distribution [33], reaching the same result from Zhang and Zhang's work [37].

The de-biased lasso is feasible when there is a good approximation of $\widetilde{G}$. To do so, a method called the lasso for node-wise regression has been suggested. Details regarding the node-wise regression and the theoretical results of asymptotic normality

for de-biased lasso can be found in [33]. Note that the residual term R is asymptotically negligible under some additional sparsity conditions when approximating $\widetilde{G}$.

It is shown that the de-biased lasso is very effective in making statistical inference about a low dimensional parameter when the sparsity assumption holds [33,37]. In Sect. 5, we investigate whether or not the de-biased lasso can outperform the lasso and other regularisation methods when increasing the data dimension in sparse situations.

3 Curse of Dimensionality with a Fixed-Effect Gaussian Design

In this section, we demonstrate, via simulations, that how the curse of dimensionality could yield undesirable and inconsistent feature selection in high dimensional regression. We generated 100 datasets from the high dimensional linear regression model (1) with standard Gaussian noise. We considered a range of different values from 100 to 5000 for the dimension p, and for each value of p we generated the design matrix $X \in \mathbb{R}^{150\times p}$ from the standard Gaussian distribution. Also for the true vector of coefficients $\beta_0 \in \mathbb{R}^p$, we generated its first 100 entries from the standard Gaussian distribution and set all the rest to 0. We fitted the lasso with L_1 penalty to each generated dataset with 10-fold cross validation using the **glmnet** package in R [11].

The results, presented in Fig. 1, show that the average identification rate tends to decrease when the dimension p increases. In other words, the number of selected variables that are correctly identified over the number of selected variables decreases as p gets larger. Since statistical inference is not feasible with lasso regression [21] as already discussed, one may rely on the fitted model which associates with the smallest prediction mean square error (MSE). From Figure 1, it can be seen that the proportion

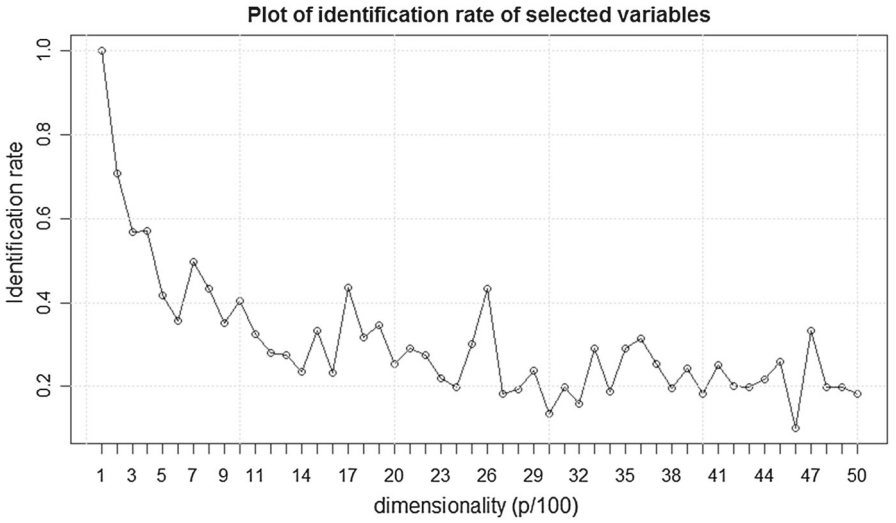


Fig. 1 Lasso becomes less effective in feature selection when p increases

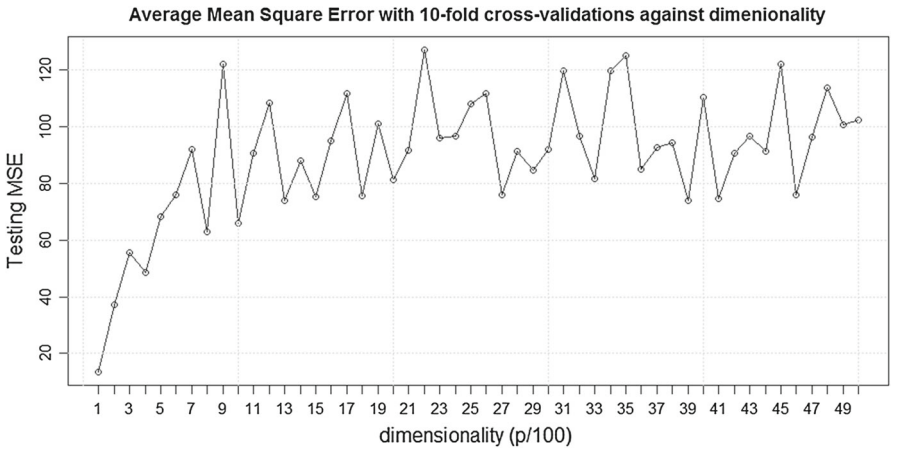


Fig. 2 The lasso regression prediction error increases with p

Table 1 The number of selected variables by lasso for different values of p in the first 10 trials

p	Trial									
	1	2	3	4	5	6	7	8	9	10
100	64	71	59	67	73	68	72	72	67	72
200	88	92	104	59	73	109	73	92	51	80
300	51	61	74	70	93	30	68	54	46	23
400	45	54	32	59	53	28	49	40	34	37
500	105	41	85	94	60	83	91	82	122	93
1000	45	44	0	36	35	29	58	19	9	10
2000	57	104	61	76	75	83	55	52	47	89
3000	1	0	66	0	107	6	8	48	115	4
4000	44	85	80	27	30	1	1	3	3	22
5000	100	48	90	53	36	6	27	29	61	34

of non-zero variables correctly identified by lasso regression over those which are selected essentially decreases from 100% to 18.2% when moving from $p = 100$ towards $p = 5000$. This is also reflected by the increasing MSE as shown in Fig. 2. Table 1 provides more details regarding the first 10 trials where we can see that the number of variables selected by lasso regression seems to be consistent from $p = 100$ to $p = 500$, but deteriorates as p becomes much larger. All these suggest feature selection inconsistency of lasso regression in high dimensional situations with very large p.

The choice of λ during cross-validation is crucial since λ governs the threshold value below which the estimated coefficients are forced to be 0. In our simulations, we used the default cross-validation values in the **glmnet** package. Figure 3 shows the MSE for different values of λ from a 10-fold cross-validation in a single trial when $p = 5000$. Here the grid used to determine an optimal choice of λ is

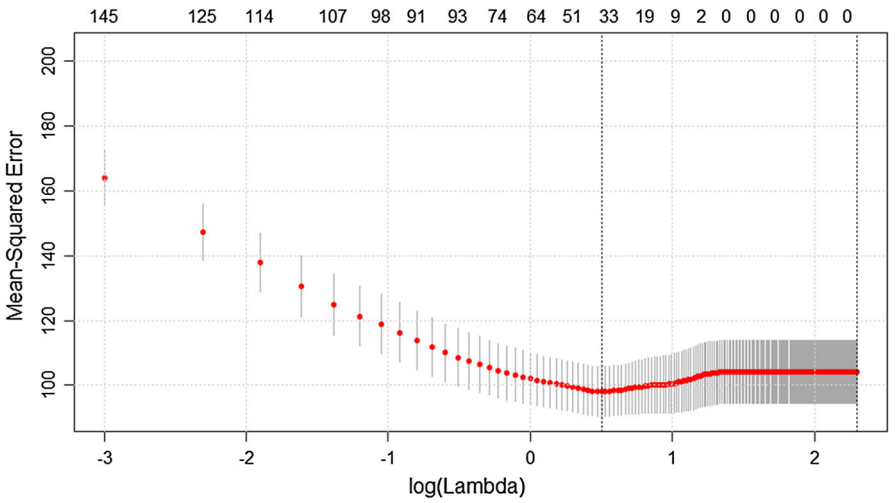


Fig. 3 Cross validation plot for optimising λ in the lasso regression, with associated number of variables selected on the top with each candidate of λ. The left vertical dashed line refers to the candidate associated with the minimum MSE, and the right vertical dashed line refers to the largest candidate which is within 1 standard deviation away from the minimum MSE

{0.05, 0.1, 0.15, . . . , 0.95, 1, 1.05, . . . , 1.95, 2, . . . , 10} instead of the default ones. In this case, the one associated with the minimum MSE is 1.65. Compared to Fig. 4 which uses the default values in **glmnet**, Fig. 3 reveals another aspect of the curse of dimensionality: the error bars are too large for every cross-validated values of λ that even 1.65 may not be a good choice for optimal λ after all, and this is mainly due to the lack of sufficient observations compared to the number of covariates or features. One may choose a wider range for the grid, however substantial standard errors would be unavoidable.

The cross-validation result presented in Fig. 3 may not alone reveal all aspects of the situation. Figure 5 presents a more general result on the cross validation from the first 20 trials with the same simulation setting and with $p = 5000$. It can be seen that there are different cross-validation patterns, all with substantial standard errors. We recall that in our example the sample size is 150 (in accordance with the common applications of high dimensional data) with the number of truly non-zero coefficients being 100. With the vast number of covariates, the lasso may still lead to inconsistent variable selection when there is a large number of truly non-zero coefficients [38].

4 Performance of Regularisation Methods in Sparse and Non-sparse High Dimensional Data

In this section, we investigate the performance of three common regularisation methods (lasso, ridge regression and elastic net) in estimation, prediction and variable selection for high dimensional data under different sparse and non-sparse situations. In particular, we study the performance of these regularisation methods when **data**

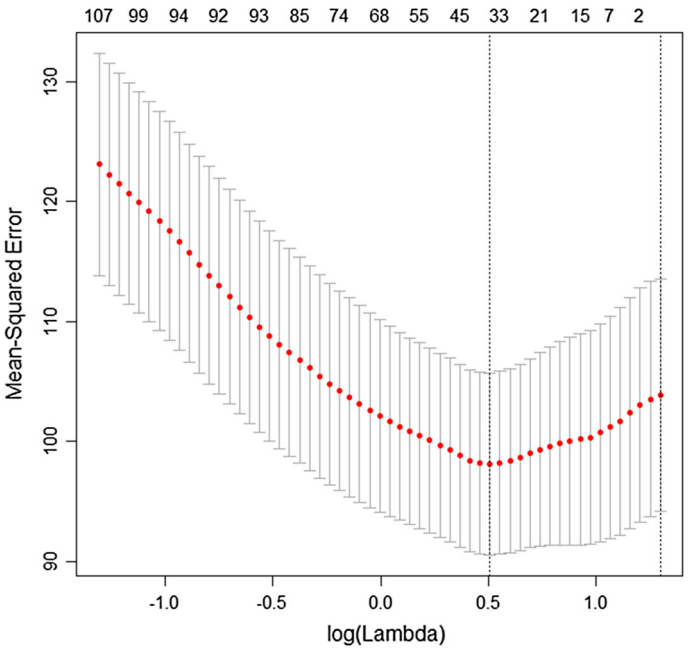


Fig. 4 A similar cross validation plot to Fig. 3 with the default grid for comparisons

correlation, **covariate location** and **effect size** are taken into account in the high dimensional linear model (1), as explained in the sequel.

We assume the true underlying model is

$$y_i = \underline{x}_i^T \beta^0 + \varepsilon_i, \quad \varepsilon_i \sim N(0, \sigma^2 = 1), \tag{5}$$

where $\beta^0 \in \mathbb{R}^p$ is the vector of true regression coefficients.
Also, we define the following notations:

- $\hat{\beta} \in \mathbb{R}^p$: estimator of coefficients in the fitted model,
- $\hat{y} = X\hat{\beta} \in \mathbb{R}^n$: vector of predicted values using the fitted model,
- S_0: active set of variables of the true underlying model,
- S_{final}: active set of the fitted model,
- $S_{\text{small}} \subset S_0$: active subset of variables with small contributions in the true underlying model,

where the active set refers to the index set representing the covariates in the regression model. We use the following performance measures to assess the accuracy of prediction, parameter estimation and identification rates for each method:

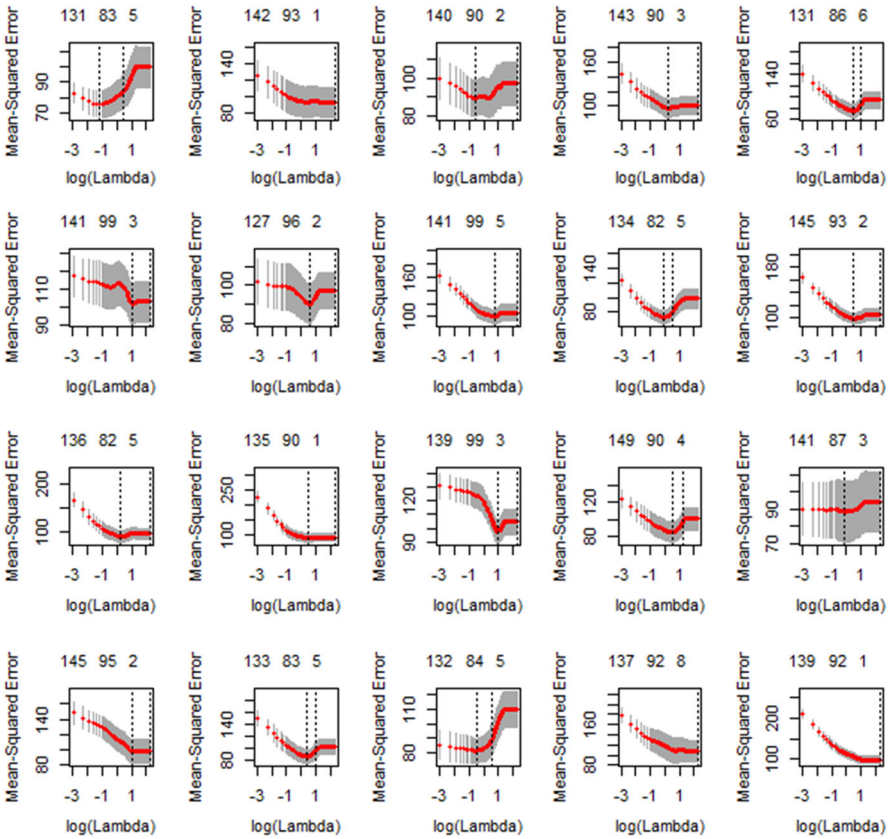


Fig. 5 Cross-validation plots of the first 20 trials with $p = 5000$

$$
\begin{aligned}
\text{Mean Square Error (MSE)} &= \frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - y_i)^2 \\
\text{Mean Absolute Bias (MAB)} &= \frac{1}{|S_{\text{final}} \cap S_0|} \sum_{j \in S_{\text{final}} \cap S_0} |\hat{\beta}_j - \beta_j^0| \\
\text{Power } (P) &= \frac{|S_{\text{final}} \cap S_0|}{|S_0|} \\
\text{Small Power } (P_{\text{small}}) &= \frac{|S_{\text{final}} \cap S_{\text{small}}|}{|S_{\text{small}}|}.
\end{aligned} \tag{6}
$$

In the simulations, we generate X and β^0 from a zero-mean multivariate Gaussian distribution with covariance matrices Σ_X and Σ_{β^0} chosen under different scenarios. Σ_{β^0} is chosen as the identity matrix when generating β^0, however we change the diagonal entries from 1 to 0.1 when coefficients of small effects are considered in the simulations. We use the identity matrix for Σ_X when no data correlation is present, and we consider $\Sigma_X = V_3$ when inducing correlated data in X, where V_i is defined, for any positive integer $i \in \mathbb{Z}^+$, as follows

$$\Sigma_X = V_i := \left[\begin{array}{cc|c} A & & 0 \\ & \ddots & \\ 0 & & A \\ \hline & 0 & I \end{array}\right], \quad \text{with } A = \overbrace{\begin{bmatrix} 1 & 0.8 & \dots & \dots & 0.8 \\ 0.8 & 1 & 0.8 & \dots & \vdots \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ \vdots & \vdots & \ddots & \ddots & 0.8 \\ 0.8 & \dots & \dots & 0.8 & 1 \end{bmatrix}}^{\text{300 columns, same for rows}},$$

in which the top-left block of V_i contains i matrices of A, which are aligned diagonally. Also, the bottom-right block is an identity matrix to fulfil the required dimension p.

In a sparse situation, where the underlying data structure is truly sparse, we choose $n = 150$, $p = 10{,}000$, and $p^* = 200$. Recall that $p^* > n$ means it is impossible to identify all of the important features without over-fitting. Given p can be much larger with the same n in practice [2], such identification may not be possible even in a sparse situation, thus β_j^0 is set to be 0 for any $j = n+1, n+2, \dots, p$ unless stated otherwise. In a non-sparse situation, we change p^* to 1000. In this case, we either use $\Sigma_X = V_5$ or $\Sigma_X = V_{20}$ to account for data correlation.

Each simulation is repeated 100 times and the average values are calculated for each performance measure in (6). We use the R-package **glmnet** for implementing the regularisation methods considered here.

We also include the principal component regression (PCR) [10,19] in our comparisons. PCR is becoming a popular technique in different fields, especially in bioinformatics [22]. By performing principal component analysis (PCA) on X whose columns are centred, one obtains an orthonormal basis with k principal components (called the loading matrix $B \in \mathbb{R}^{p\times k}$), which are used to transform X to a new design matrix $X' = XB$, $X' \in \mathbb{R}^{n\times k}$. One can then perform the OLS on X' by considering

$$y = X'\gamma + \varepsilon,$$

where y is mean-centred and $\gamma \in \mathbb{R}^{k\times 1}$ is the vector of unknown transformed parameters to be estimated. With the estimator $\hat{\gamma} = (X'^T X')^{-1} X'^T y$, one can find $\hat{\beta}$ by reverting the transformation as follows

$$\hat{\beta} = B\hat{\gamma}.$$

In high dimensional situations, when $p > n$, there can be at most $n-1$ principal components, which means the OLS with X' does not face the ill-posed problem. PCR exists for a long time [19,20], but it did not attract much attention before partly because it could be seen as a hard thresholding version of ridge regression from the perspective of singular value decomposition [10]. The magnitude of eigenvalues associated to each principal component corresponds to the amount of information not redundant from X. To find the optimal number of principal components, we perform 10-fold cross-validation on each candidate by successive inclusion of principal components in decreasing order of their associated eigenvalues. We use the R-package **pls** [34] for implementation of PCR in our simulations.

4.1 Data Correlation

The complete simulation results for all the four methods under the sparse and non-sparse situations are reported in Tables 2 and 3 below. Regarding the prediction performance, a part of the simulation results presented in Fig. 6 shows that, when important covariates are associated with correlated data, the prediction performance of lasso and elastic net are better than both the ridge regression and PCR under the sparse situation. Another part of the simulation results shown in Fig. 7 suggests that, for the non-sparse situation, the prediction performance of the lasso and elastic net is very similar to the ridge regression, however PCR outperforms all of these methods.

Regarding the parameter estimation accuracy, the simulation results in Figs. 6 and 7 indicate that parameter estimation by ridge regression and PCR are largely unaffected by correlated data, with both having smaller average MAB compared to the lasso and elastic net, and this performance is mainly because of their dense solutions.

Regarding the variable selection performance, Fig. 7 shows that elastic net performs better than lasso in the case of correlated data, which can be justified by the presence of the grouping effect. Note that the data correlation associated with nuisance and less important variables seems to have little effect on our results compared to the data correlation associated with important variables.

Without data correlation, lasso and elastic net have prediction performances similar to ridge regression and PCR under the sparse situation (see the results for case 1 in Table 2). This is probably because of the identification of important covariates being limited by sample size and high dimensionality, causing difficulty for the lasso and elastic net to outperform the ridge regression and PCR. Under the non-sparse situation, lasso and elastic net performed even worse in prediction compared to ridge regression and PCR (see the results for case 1 in Table 3).

Overall, when important covariates are associated with correlated data, our results showed that the prediction performance is improved across all these four methods under both sparse and non-sparse situations, and that the prediction performance flipped to favour the lasso and elastic net over the ridge regression and PCR.

4.2 Covariate Location

Regarding the effects of the covariate location, we find, from the simulation results, that important variables being more scattered among groups of correlated data tend to result in better prediction performances. Such observation becomes more obvious under the non-sparse situation. With the same data correlation setting, all the methods performed better with 2 clusters of size 500 in Fig. 8, than with 5 clusters of important variables of size 200 as shown in Fig. 7. Since lasso tends to randomly select covariates in a group of correlated data, we expect that the lasso is less likely to select nuisance covariates when most of them are important in such group, thus improving prediction and variable selection performances. In terms of estimation accuracy, ridge regression and PCR were largely unaffected as expected, while lasso and elastic net had varying results. Therefore, the condition of covariate location helping prediction performance does not seem to necessarily reflect in the parameter estimation.

Table 2 Complete results of simulation I

Case	Indices j: β_j^0 is generated from N(0,1)	Indices j: β_j^0 is generated from N(0,0.1)	Choice of Σ_X
Set-up for simulation study I with a sparse situation			
1	1–200	NA	$I_{p\times p}$
2	101–200, 701–800	NA	V_3
3	101–170, 351–420, 701–760	NA	V_3
4	101–200	201–300	$I_{p\times p}$
5	251–350	551–650	V_3
6	291–310	401–580	V_3
7	201–380	591–610	V_3

Case	Lasso	Ridge	Elastic Net (0.5)	PCR
Average MSE summary of simulation I				
1	203.0375	193.0763	200.8516	194.7625
2	46.4572	58.9701	44.8423	53.0762
3	49.8918	61.0208	49.3757	53.6081
4	92.7584	92.2995	91.3694	92.7680
5	21.6402	27.3076	21.5200	25.1636
6	4.3967	6.6928	4.6757	5.3609
7	38.9556	44.9188	37.9353	40.4747
Average MAB summary of simulation I				
1	1.1797	0.7910	1.2846	0.7910
2	0.9208	0.8114	0.8986	0.8118
3	0.8537	0.7918	0.8890	0.7922
4	1.3710	0.4368	1.3741	0.4368
5	0.8510	0.4405	0.8764	0.4409
6	0.6222	0.1513	0.6144	0.1514
7	0.8116	0.7217	0.9442	0.7221

Case	Lasso	Elastic Net (0.5)
Average Power (P) summary of simulation I		
1	0.0112	0.0153
2	0.0633	0.0880
3	0.0598	0.0835
4	0.0143	0.0187
5	0.0493	0.0705
6	0.0363	0.0528
7	0.0513	0.0768
Average Small Power (P_{small}) summary of simulation I		
4	0.0023	0.0033
5	0.0133	0.0200
6	0.0150	0.0298
7	0.0083	0.0133

Table 3 Complete results of simulation III

Case	Indices j: β_j^0 is generated from N(0,1)	Indices j: β_j^0 is generated from N(0,0.1)	Choice of Σ_X
Set-up for simulation III with a non-sparse situation			
1	1–1000	NA	$I_{p\times p}$
2	51–250, 351–550, 651–850, 951–1150, 1251–1450	NA	V_5
3	1–500, 801–1300	NA	V_5
4	1–500, 801–1300	NA	V_{20}
5	1–500	501–1000	$I_{p\times p}$
6	1–500	501–1000	V_{20}
7	101–150	401–1350	V_{20}
8	101–1050	1401–1450	V_{20}

Case	Lasso	Ridge	Elastic Net (0.5)	PCR
Average MSE summary of simulation III				
1	1107.7110	1057.6660	1091.0470	1069.1520
2	294.8247	294.3531	290.2807	278.7965
3	259.0461	259.3874	255.1258	242.4962
4	265.4878	262.2610	256.5398	249.1404
5	522.7113	507.2834	517.5554	511.7361
6	122.7950	138.4813	121.0770	132.2491
7	14.9461	18.3394	14.9496	16.2734
8	230.4608	240.7334	226.3334	235.7741
Average MAB summary of simulation III				
1	1.3913	0.7954	0.9473	0.7957
2	1.1660	0.8035	0.9259	0.8039
3	1.1698	0.7996	0.8868	0.8001
4	1.1487	0.7958	0.9016	0.7960
5	0.7533	0.4345	0.7058	0.4346
6	0.9142	0.4385	0.7920	0.4388
7	0.4784	0.1156	0.4433	0.1157
8	1.1093	0.7633	0.8896	0.7639

Case	Lasso	Elastic Net (0.5)
Average Power (P) summary of simulation III		
1	0.0022	0.0021
2	0.0228	0.0339
3	0.0240	0.0372
4	0.0272	0.0399
5	0.0018	0.0035
6	0.0244	0.0360
7	0.0145	0.0202
8	0.0272	0.0401

Table 3 continued

Case	Lasso	Elastic Net (0.5)
Average Small Power (P_{small}) summary of simulation III		
Case	lasso	Elastic Net (0.5)
5	0.0015	0.0025
6	0.0076	0.0121
7	0.0092	0.0134
8	0.0013	0.0020

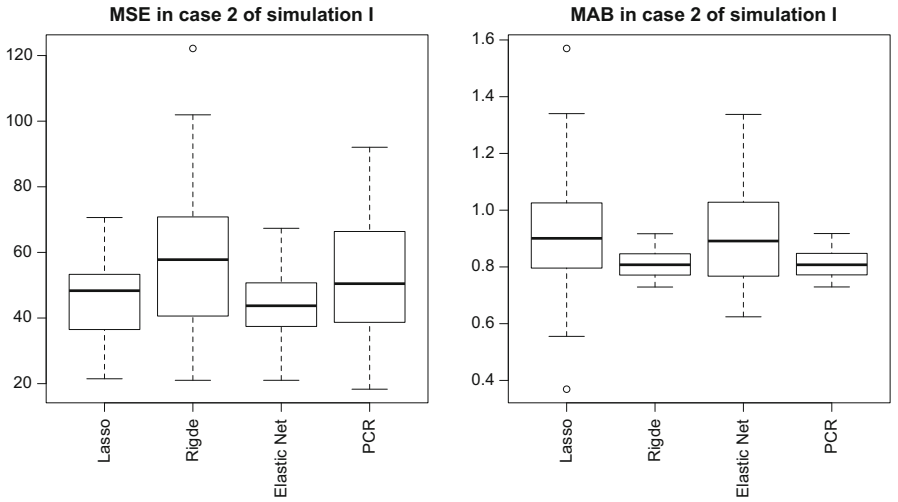


Fig. 6 The average MSE and MAB with the data correlation associated with important variables under a sparse situation

4.3 Effect Size

Given the same number of important covariates, our simulation results (see cases 1 and 4 in Table 2 and cases 1 and 5 in Table 3) suggest that having a smaller overall effect size helps prediction and parameter estimation performances across all the methods. This is reasonable since the magnitude of errors is smaller in exchange of harder detection of covariates, having small contributions to the predictions.

With data correlation, our results also reveal that the overall effect size could alter our perception of underlying data structures in the non-sparse situation. Figure 9 shows the performance bar-plots for all the four methods when there were 1000 important covariates, 950 of which belonging to small effect size. Compared to Figs. 7 and 8 that both of which had 1000 important covariates of similar effect sizes, Fig. 9 indicates that the lasso and elastic net tend to perform better than the ridge regression and PCR in terms of prediction accuracy in this situation. This is probably because selecting some of those 50 important features associated with large effects is sufficient to explain the majority of the effects behind, which masks those associated with small effects. Other

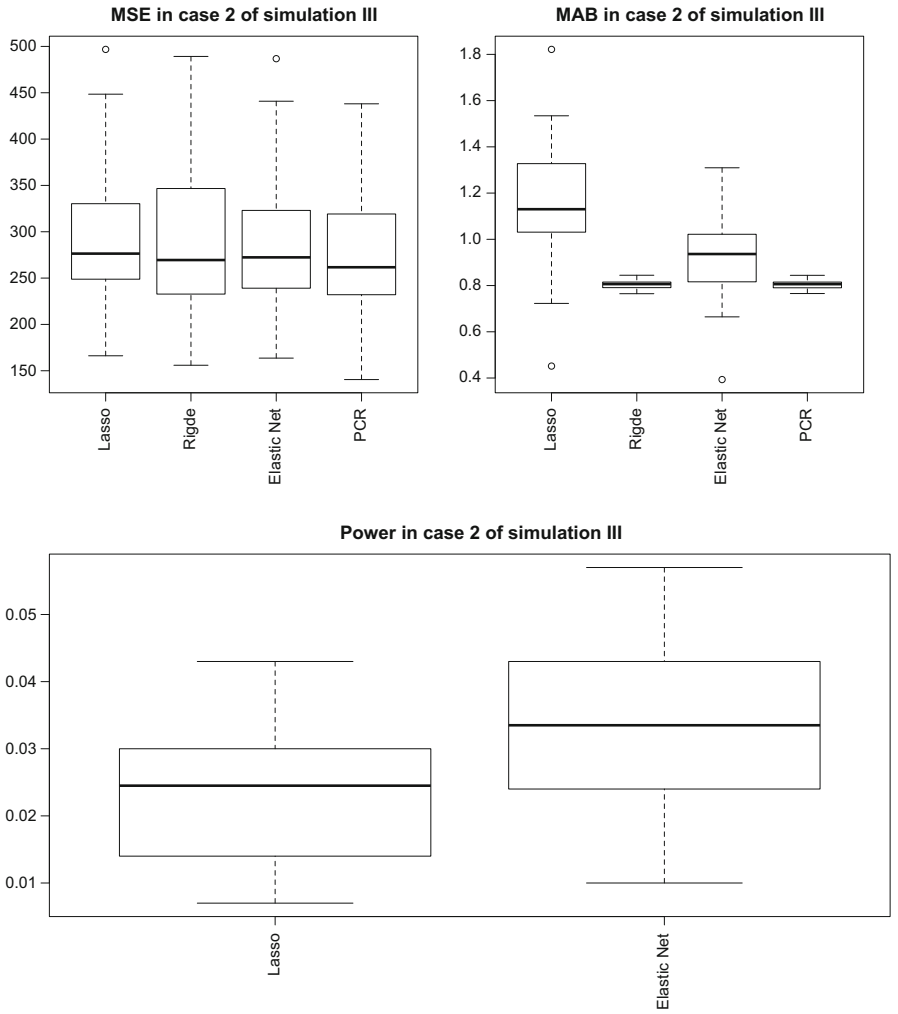


Fig. 7 The average MSE, MAB and power with the data correlation associated with important variables under a non-sparse situation

than the sparsity level of important covariates, overall covariate effect size seems to also change the indication of whether an underlying data structure is sparse via observing prediction performances, especially in a non-sparse situation.

5 Performance of the De-Biased Lasso

Similar to the lasso, sparsity assumptions play a major role in justifying the use of de-biased lasso. In this section, we evaluate the performance of the de-biased lasso in prediction, parameter estimation and variable selection, and compare the results with

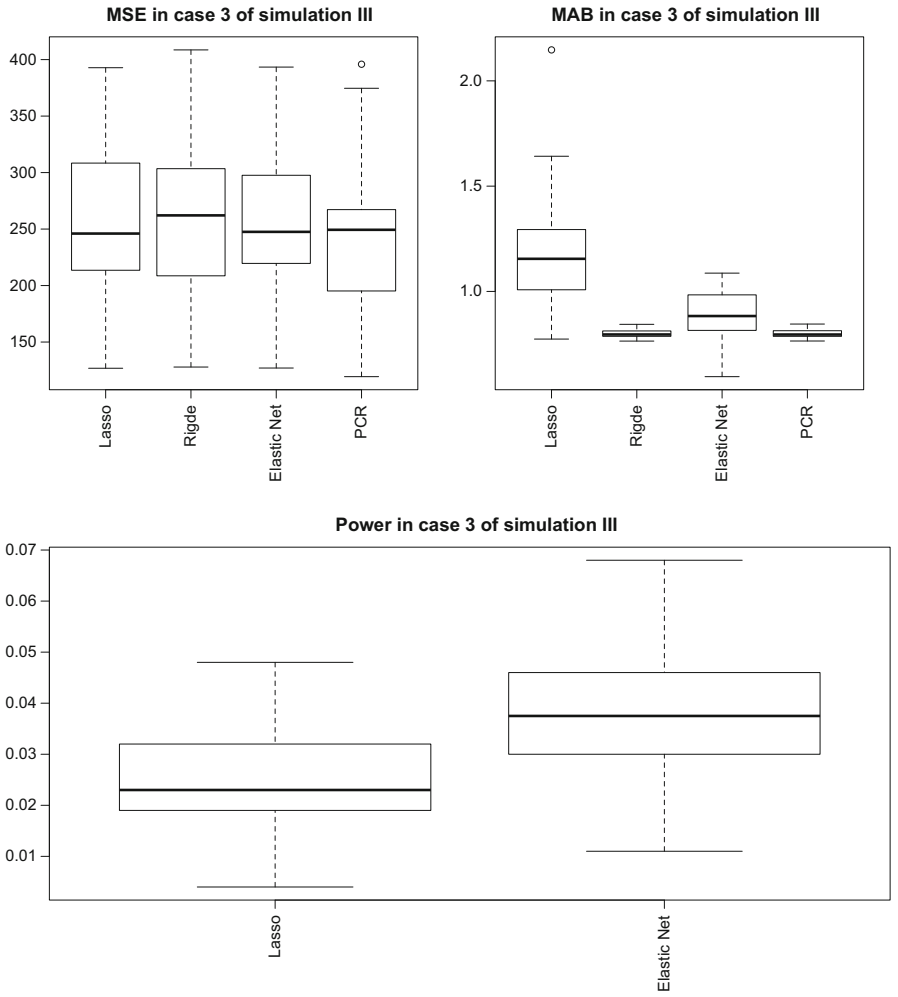


Fig. 8 The average MSE, MAB and power with important variables being more concentrated across the groups of correlated data under a non-sparse situation

the other methods considered in the previous section. We are particularly interested in understanding how this recently developed method performs when the data dimension p increases, so we can provide a rough idea of its practicality to emerging challenges in big data analysis.

In the simulations, we again focus on high dimensional situations with the effects of **data correlation**, **covariate location** and **effect size** being considered. We use the sample size $n = 40$ and let the maximum dimension p to be 600 here. Similar to the previous simulations, we repeat each simulation case 100 times and calculate the average values for each performance measure in (6). We should mention that for calculating P and P_{small} both the lasso and elastic net are based on their nature of variable selection, but the de-biased lasso is based on point-wise statistical inference

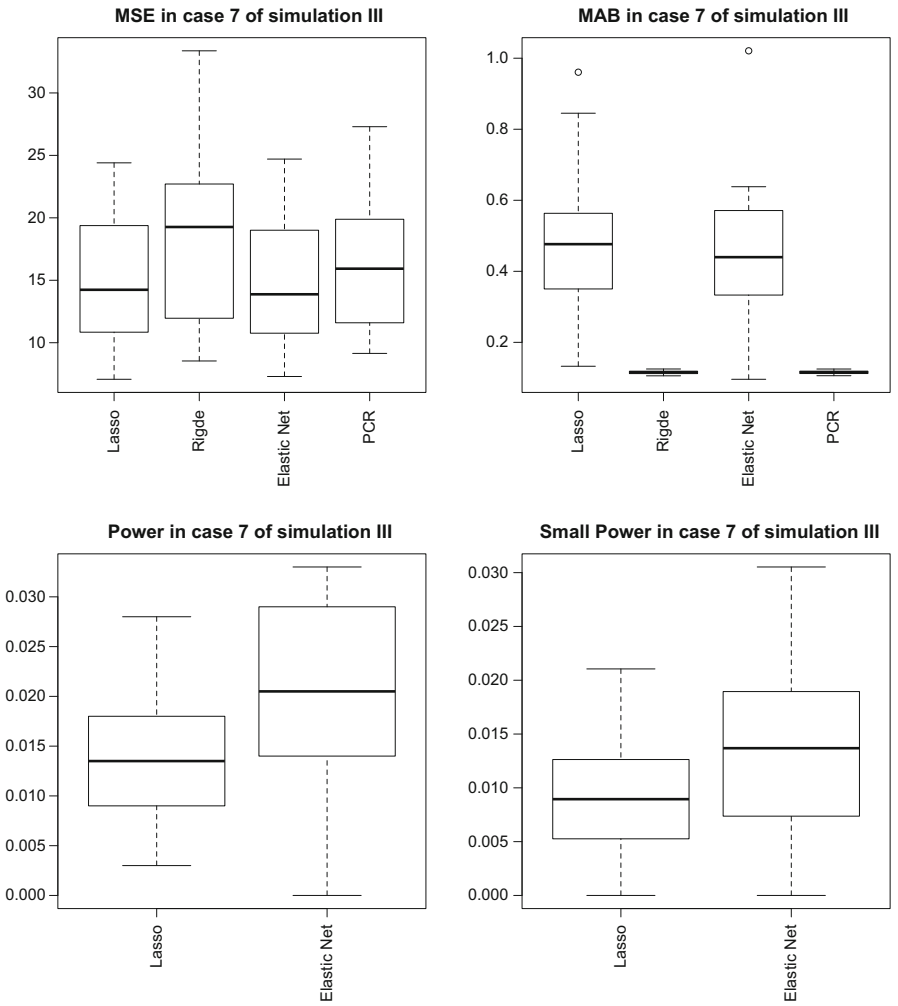


Fig. 9 The average MSE, MAB, power and small power for the case with the majority of important covariates having small contributions among groups of correlated data under a non-sparse situation

[33,37]. This implies that the performance of de-biased lasso on P and P_{small} can encounter multiple hypothesis testing issue on top of curse of dimensionality. To account for this, we corrected the associated p-values by controlling the family-wise error rate of the tests via the Bonferroni-Holm approach [16]. To induce correlated data, we consider $\Sigma_X = V_2$. We use the R-package **hdi** [7] to implement the de-biased lasso to each generated data set.

The complete simulation results are given in Table 4 below. A part of the results, presented in Fig. 10, show that in a low dimensional sparse situation with uncorrelated data, the de-biased lasso outperforms all the other methods in terms of prediction and parameter estimation. However, the variable selection performance of the de-

Table 4 Complete results of simulation IV

Case	Dimension p	Number of truly non-zero coefficients p^*	Indices $j : \beta_j^0$ is generated from N(0,1)	Indices $j : \beta_j^0$ is generated from N(0,0.1)	Choice of Σ_X
Set-up information for simulation IV					
1	50	5	1–5	NA	$I_{p\times p}$
2	100	10	1–10	NA	$I_{p\times p}$
3	600	60	1–60	NA	$I_{p\times p}$
4	600	60	1–60	NA	V_2
5	600	60	1–50	401–410	V_2

Case	lasso	Ridge	Elastic Net (0.5)	PCR	De-biased lasso
Average MSE summary of Simulation IV					
1	1.0286	3.8180	1.1865	2.9341	0.8026
2	4.9766	9.9611	5.1586	9.7301	4.7964
3	60.0761	55.1843	58.4472	59.0408	59.1018
4	16.6779	17.4790	17.5231	14.5093	42.8449
5	11.6343	9.4145	9.8804	8.2314	23.1490
Average MAB summary of Simulation IV					
1	0.3150	0.7965	0.3936	0.4798	0.0842
2	0.5335	0.7704	0.5786	0.6290	0.2760
3	0.8880	0.7709	1.0084	0.7655	NA

Table 4 continued

Case	lasso	Ridge	Elastic Net (0.5)	PCR	De-biased lasso
4	0.8551	0.7785	1.0063	0.7709	2.7113
5	1.2956	0.7114	1.0709	0.7036	1.9395

Case	lasso	Elastic Net (0.5)	De-biased lasso
Average power (P) summary of Simulation IV			
1	0.74	0.78	0.72
2	0.52	0.48	0.24
3	0.0233	0.03	0
4	0.0533	0.0833	0.0017
5	0.0383	0.0650	0.0083
Average small power (P_{small}) summary of Simulation IV			
5	0	0.01	0

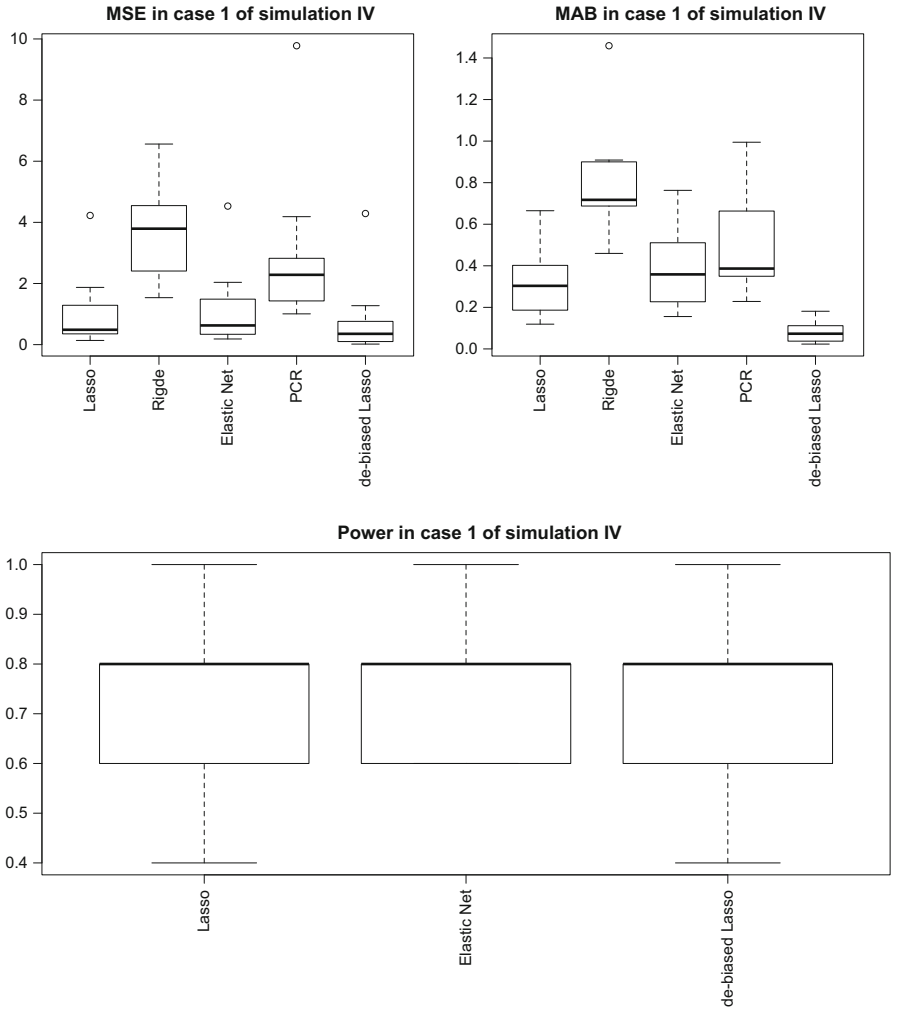


Fig. 10 The average MSE, MAB and power in sparse high dimensional data with dimension $p = 50$

biased lasso is very similar to the lasso and elastic net. The results shown in Fig. 11 suggest that the prediction by de-biased lasso could still be as good as the lasso and elastic net in a sparse situation when the dimension p increases, however the de-biased lasso no longer identifies any important covariates, thus its performance in parameter estimation cannot be assessed in the case when the dimension p is large. The results in Fig. 12 show that inducing correlated data seems to help de-biased lasso identify important covariates, however its performance in prediction and parameter estimation is no longer comparable to the other methods in the case of correlated data. The unsatisfactory variable selection performance of the de-biased lasso is probably due to many hypothesis tests as mentioned above, and its poor performance in prediction and parameter estimation in the case of correlated data could be due to the multicollinearity issues causing spurious test results.

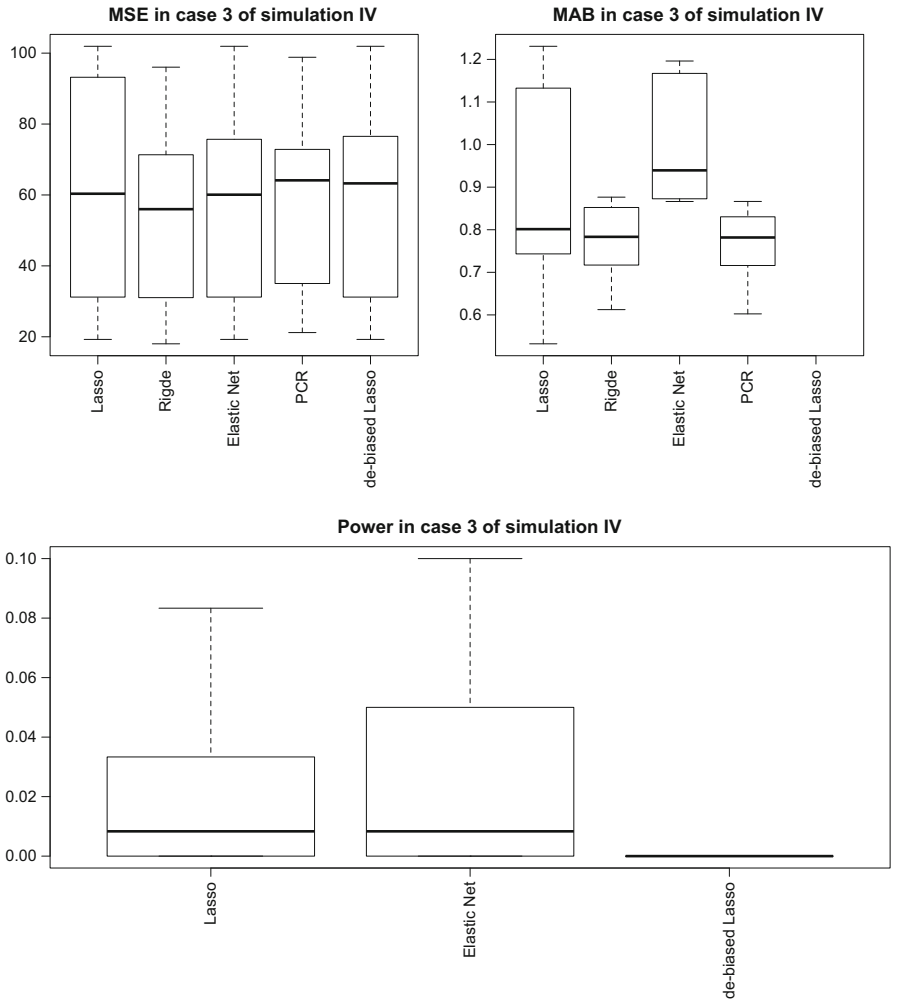


Fig. 11 The average MSE, MAB and power in sparse high dimensional data with dimension $p = 600$

6 Real Data Example

In this section, we use a real data example to compare the performance of the different regularisation methods in real-world applications. We consider the riboflavin data obtained from a high-throughput genomic study concerning the riboflavin (vitamin B_2) production rate. This data set was made publicly available by Bühlmann et al. [3], and contains $n = 71$ samples and $p = 4088$ covariates corresponding to $p = 4088$ genes. For each sample, there is a real-valued response variable indicating the logarithm of the riboflavin production rate along with the logarithm of the expression level of the $p = 4088$ genes as the covariates. Further details regarding the dataset and its availability can be found in [7,17].

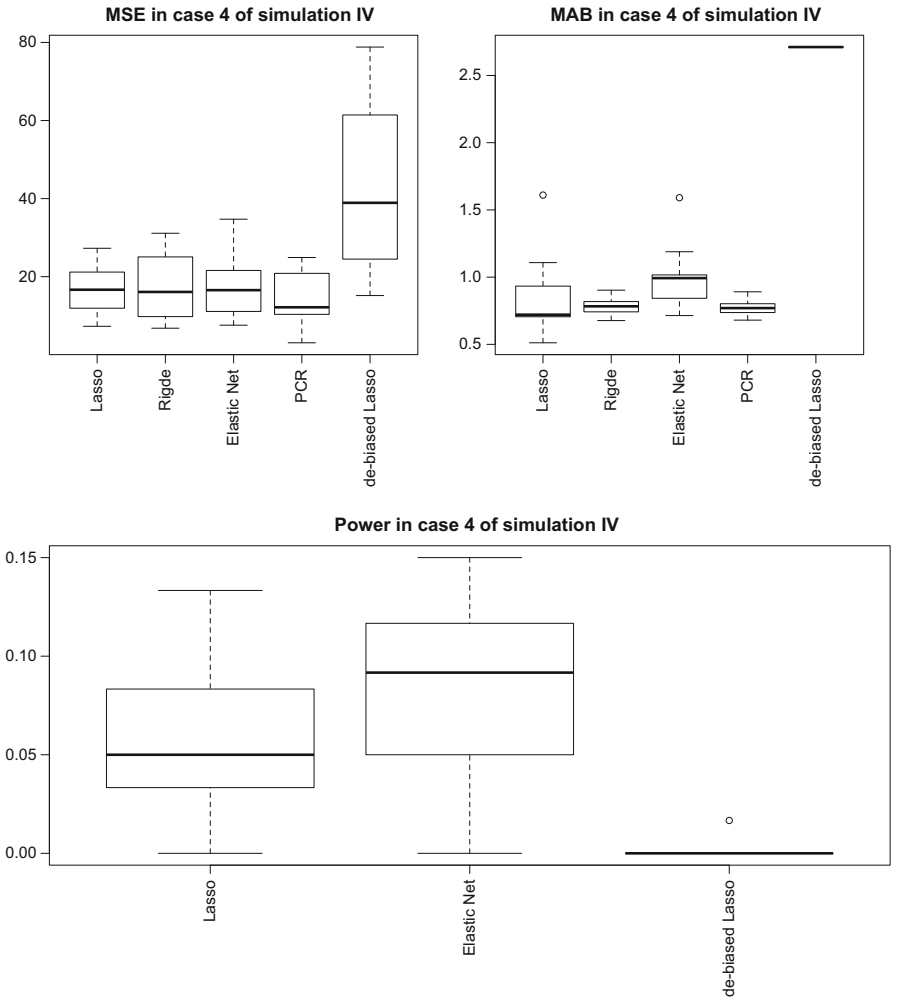


Fig. 12 The average MSE, MAB and power in sparse high dimensional data with dimension $p = 600$ and with the presence of correlated data

Table 5 The prediction results from applying all the methods to the riboflavin data

Lasso	Ridge regression	Elastic net	PCR	De-biased lasso
The average MSE on the test data				
0.2946	0.3953	0.3331	0.3493	0.3278

We applied the de-biased lasso and each of the other methods to the riboflavin data 100 times through different random partitions of training and testing sets, and compared their prediction performance using the average MSE as in (6). The prediction results are shown in Table 5. The results indicate that while PCR performed as good if not better than ridge regression, the lasso and elastic net had smaller average MSE

than both ridge regression and PCR. The MSE of de-biased lasso is similar to the elastic net and lasso. The potential correlation between genes helps elastic net and lasso to perform better in prediction, which is consistent with our simulation results in the previous sections. Also, according to our simulation findings in Sects. 4 and 5, it seems the underlying structure of the riboflavin dataset is sparse in the sense that among all the unknown covariates, which contribute to the production rate of vitamin B_2, only a few of them have relatively large effects. We emphasise that this does not necessarily indicate sparsity on the number of important covariates compared to the data dimension. The riboflavin dataset has been recently analysed for statistical inference purposes such as constructing confidence intervals and hypothesis tests by some researchers including [3,7,17].

7 Conclusions and Discussion

We investigated the effects of data correlation, covariate location and effect size on the performance of regularisation methods such as lasso, elastic net and ridge regression when analysing high dimensional data. We particularly evaluated how prediction, parameter estimation and variable selection by these methods are affected under those conditions. We also studied the performance of the recently developed de-biased lasso under such conditions, and furthermore included the PCR in our simulations for comparison purposes. We considered different sparse and non-sparse situations in our simulation studies. The main findings of the simulation results and real data analysis are summarised below:

- When important covariates are associated with correlated variables, the simulation results showed that the prediction performance improves across all the methods considered in the simulations, for both sparse and non-sparse high dimensional data. The prediction performance flipped to favour the lasso and elastic net over the ridge regression and PCR.
- When the correlated variables are associated with nuisance and less important variables, we observed that the prediction performance is generally unaffected across all the methods compared to the situation when the data correlation is associated with important variables.
- In the presence of correlated variables, the parameter estimation performance of the ridge regression, elastic net and PCR was not affected, but the lasso showed a poorer parameter estimation when moving from sparse data to non-sparse data.
- The variable selection performance of the elastic net was better than the lasso in the presence of correlated data.
- Regarding the effects of the covariate location, we found that important variables being more scattered among groups of correlated data tend to result in better prediction performances. Such behaviour was more obvious for non-sparse data. The lasso tends to randomly select covariates in a group of correlated data, so it is less likely to select nuisance covariates when most of them are important in such group, thus improving prediction and variable selection performances.

- Unlike in prediction and variable selection, the impact of covariate location was very small on the parameter estimation performance across all the methods.
- Given the same number of important covariates, the simulation results showed that having a smaller overall effect size helps the prediction and parameter estimation performances across all the methods. The simulation results indicated that the lasso and elastic net tend to perform better than the ridge regression and PCR in terms of prediction accuracy in such situation. In the presence of data correlation, the overall effect size could change our indication of whether an underlying data structure is sparse via observing prediction performances, especially in the non-sparse situations.
- For the de-biased lasso, the simulation results showed that the de-biased lasso outperforms all the other methods in terms of prediction and parameter estimation in low dimensional sparse situations with uncorrelated data. When the data dimension p increases, the prediction by de-biased lasso is as good as the lasso and elastic net, however the de-biased lasso no longer identifies any important covariates when the dimension p is very large. The results also showed that inducing correlated data seems to help de-biased lasso identify important covariates when p is very large, however its performance in prediction and parameter estimation is no longer comparable to the other methods in the presence of correlated data.

It should be pointed out that we also included the adaptive lasso [39] in our simulation comparisons, however because the results were very similar to the lasso we did not report them in the simulation section.

We also observed that the curse of dimensionality can yield inconsistent and undesirable feature selection in high dimensional regression. The choice of shrinkage parameter λ during the cross-validation process was found to be crucial. For high dimensional data, the error bars were too large for every cross-validated value of λ and it was mainly due to the lack of sufficient observations compared to the number of covariates ($p \gg n$).

Finally, the de-biased lasso can be used in a similar fashion as the OLS, but in ill-posed low dimensional problems. It therefore suffers from multicollinearity as well as the issue of too many hypothesis tests in high dimensional data [3,7,33]. With many procedures available to tackle issues from multiple hypothesis testing, a more accurate estimation procedure would be helpful when applying the de-biased lasso to high dimensional data. It will be very useful to conduct research on how the de-biased lasso combined with bootstrap [8] performs in high dimensional data under the above three conditions.

References

1. Ayers KL, Cordell HJ (2010) Snp selection in genome-wide and candidate gene studies via penalized logistic regression. Genet Epidemiol 34(8):879–891
2. Bühlmann P (2017) High-dimensional statistics, with applications to genome-wide association studies. EMS Surv Math Sci 4(1):45–75
3. Bühlmann P, Kalisch M, Meier L (2014) High-dimensional statistics with a view toward applications in biology. Ann Rev Stat Appl 1:255–278
4. Bühlmann P et al (2013) Statistical significance in high-dimensional linear models. Bernoulli 19(4):1212–1242
5. Cantor RM, Lange K, Sinsheimer JS (2010) Prioritizing gwas results: a review of statistical methods and recommendations for their application. Am J Hum Genet 86(1):6–22
6. Dalalyan AS, Hebiri M, Lederer J (2017) On the prediction performance of the lasso. Bernoulli 23(1):552–581
7. Dezeure R, Bühlmann P, Meier L, Meinshausen N et al (2015) High-dimensional inference: confidence intervals, p-values and r-software hdi. Stat Sci 30(4):533–558
8. Dezeure R, Bühlmann P, Zhang CH (2017) High-dimensional simultaneous inference with the bootstrap. TEST 26(4):685–719
9. Fan J, Li R (2001) Variable selection via nonconcave penalized likelihood and its oracle properties. J Am Stat Assoc 96(456):1348–1360
10. Friedman J, Hastie T, Tibshirani R (2001) The elements of statistical learning, vol 1. Springer series in statistics. Springer, New York
11. Friedman J, Hastie T, Tibshirani R (2010) Regularization paths for generalized linear models via coordinate descent. J Stat Softw 33(1):1
12. García C, García J, López Martín M, Salmerón R (2015) Collinearity: Revisiting the variance inflation factor in ridge regression. J Appl Stat 42(3):648–661
13. Guo Y, Hastie T, Tibshirani R (2006) Regularized linear discriminant analysis and its application in microarrays. Biostatistics 8(1):86–100
14. Hastie T, Tibshirani R, Wainwright M (2015) Statistical learning with sparsity: the lasso and generalizations. CRC Press, Boca Raton
15. Hoerl AE, Kennard RW (1970) Ridge regression: applications to nonorthogonal problems. Technometrics 12(1):69–82
16. Holm S (1979) A simple sequentially rejective multiple test procedure. Scand J Stat 6:65–70
17. Javanmard A, Montanari A (2014) Confidence intervals and hypothesis testing for high-dimensional regression. J Mach Learn Res 15(1):2869–2909
18. Jia J, Yu B (2010) On model selection consistency of the elastic net when p >> n. Stat Sinica 20:595–611
19. Jolliffe IT (1982) A note on the use of principal components in regression. Appl Stat 31:300–303
20. Kendall M (1957) A course in multivariate statistics. Griffin, London
21. Knight K, Fu W (2000) Asymptotics for lasso-type estimators. Ann Stat 28:1356–1378
22. Ma S, Dai Y (2011) Principal component analysis based methods in bioinformatics studies. Brief Bioinform 12(6):714–722
23. Malo N, Libiger O, Schork NJ (2008) Accommodating linkage disequilibrium in genetic-association analyses via ridge regression. Am J Hum Genet 82(2):375–385
24. Marafino BJ, Boscardin WJ, Dudley RA (2015) Efficient and sparse feature selection for biomedical text classification via the elastic net: application to icu risk stratification from nursing notes. J Biomed Inform 54:114–120
25. Meinshausen N (2007) Relaxed lasso. Comput Stat Data Anal 52(1):374–393
26. Nie F, Huang H, Cai X, Ding CH (2010) Efficient and robust feature selection via joint l2, 1-norms minimization. In: Advances in neural information processing systems, pp 1813–1821
27. Obenchain R (1977) Classical f-tests and confidence regions for ridge regression. Technometrics 19(4):429–439
28. Park T, Casella G (2008) The Bayesian lasso. J Am Stat Assoc 103(482):681–686
29. Ročková V, George EI (2018) The spike-and-slab lasso. J Am Stat Assoc 113(521):431–444
30. Ryali S, Chen T, Supekar K, Menon V (2012) Estimation of functional connectivity in fMRI data using stability selection-based sparse partial correlation with elastic net penalty. NeuroImage 59(4):3852–3861

31. Tibshirani R (1996) Regression shrinkage and selection via the lasso. J R Stat Soc Ser B 58:267–288
32. Tibshirani R (2011) Regression shrinkage and selection via the lasso: a retrospective. J R Stat Soc Ser B 73(3):273–282
33. Van de Geer S, Bühlmann P, Ritov Y, Dezeure R et al (2014) On asymptotically optimal confidence regions and tests for high-dimensional models. Ann Stat 42(3):1166–1202
34. Mevik B-H, Wehrens R (2007) The pls package: principal component and partial least squares regression in R. J Stat Soft 18:1–24
35. Wu TT, Chen YF, Hastie T, Sobel E, Lange K (2009) Genome-wide association analysis by lasso penalized logistic regression. Bioinformatics 25(6):714–721
36. Yuan M, Lin Y (2006) Model selection and estimation in regression with grouped variables. J R Stat Soc Ser B 68(1):49–67
37. Zhang CH, Zhang SS (2014) Confidence intervals for low dimensional parameters in high dimensional linear models. J R Stat Soc Ser B 76(1):217–242
38. Zhao P, Yu B (2006) On model selection consistency of lasso. J Mach Learn Res 7(Nov):2541–2563
39. Zou H (2006) The adaptive lasso and its oracle properties. J Am Stat Assoc 101(476):1418–1429
40. Zou H, Hastie T (2005) Regularization and variable selection via the elastic net. J R Stat Soc Ser B 67(2):301–320